基于超网络的微博相似度及其在微博舆情主题发现中的应用*

■ 梁晓贺 田儒雅 吴蕾 张学福
中国农业科学院农业信息研究所 北京 100081

摘 要：[目的/意义]准确地计算微博相似度可以提高微博主题挖掘效率，对舆情治理、保障信息安全具有实践意义。针对微博文本语义稀疏、高维的问题，提出一种融入微博非文本特征的超边相似度算法。[方法/过程]分析微博舆情发生机制，利用超网络模型表示微博舆情主题形成过程，通过计算各层子网相似度及各层子网对主题形成的贡献度构建超边相似度算法。[结果/结论]研究发现，论文所提出的相似度方法有助于提升微博舆情信息的主题聚类效果，特别是对于文字性表述相似程度高的微博信息，具有明显的主题区分性。

关键词：超边相似度 主题发现 超网络 微博
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.11.009

1 引言

随着Web2.0时代的到来，微博得到了蓬勃发展，普通网民、网络名人、新闻媒体和政府机构都将微博作为获取信息、发表评论的主要途径[1]。微博是一种用户产生内容(User Generated Content，UGC)模式，用户通过文字、表情符号、图片、视频和直播等多种媒体形式自由地表达对某一事件的观点和看法，这些信息通过微博用户的关注、转发、评论关系实现以点到面的快速传播，这极易形成舆情事件。进行微博舆情信息挖掘对预测未来事件、保障信息安全、监测舆情动态具有重要意义[2]。面对大规模微博文本，如何高效、准确地识别主题信息已经成为了人们研究的热点[3]。微博文本的相似度算法对于理解和分析文本起着至关重要的作用，被大量用于微博文本分类[4]、聚类[5]、用户推荐[6]等多个领域，而相似度算法的优劣决定着这些应用的性能。开展微博相似度研究不仅能为我国微博舆情监测提供理论方法支持，还为我国舆情管理提供决策支持。

微博文本具有内容短、信息描述能力弱、主题分散等特点，给微博相似度研究带来挑战。目前针对微博类短文本的相似度研究存在的主要难题是数据语义稀疏，本研究提出一种在文本分析基础上融入微博非文本特征的超边相似度算法，扩展了相似度算法分析对象，实现了微博关联关系的深层次识别，提高了微博舆情主题识别准确度。

2 相关研究

2.1 微博相似度分析

目前，研究人员已经提出了一些关于微博短文本的相似度计算方法，这些方法大致分为两大类：一是针对微博短文本内容特征，改进相似度算法。包括增加外部语料库方法，如A. Islam等[7]首先构建最长公共子序列，借助外部语料库的文本语义相关性关系计算文本的语义相似度。H. Ma等[8]通过挖掘语料库中具有共现关系和类别同向关系的频繁项集，构建特征词相似性矩阵来扩展短文本特征；挖掘短文本内容特征方法，如H. Wen等[9]通过对特征词进行词性标注和概

* 本文系中国农业科学院科技创新工程项目“科技情报分析与评估创新团队”(项目编号：CAAS-ASTIP-2016-AII)和中国农业科学院农业信息研究所基本科研业务费项目“基于加权策略的大数据微博突发舆情主题挖掘”(项目编号：JBYW-AII-2017-29)研究成果之一。
作者简介：梁晓贺(ORCID：0000-0003-2005-3401)，助理研究员，博士，E-mail：liangxiaohe@caas.cn；田儒雅(ORCID：0000-0002-9944-2081)，助理研究员，博士；吴蕾(ORCID：0000-0003-0514-2203)，助理研究员，博士；张学福(ORCID：0000-0002-9387-7527)，研究员，博士生导师。
收稿日期：2019-01-04 修回日期：2020-01-21 本文起止页码：77-86 本文责任编辑：杜杏叶

念标注,侧重对文本语义相似性的计算,黄贤英等[10]基于词形和词义构造了文本公共块,依据公共块的词项数量和组和顺序度量文本相似性。二是借助社会网络分析方法,引入微博的非文本特征进行相似度计算。其中,引入用户特征研究最为广泛,相关研究主要集中在引入用户自身背景信息构建相似度公式[12]、引入用户关系如关注与被关注关系[13]、共同邻居好友数量[13]构建相似度公式。随着研究的深入,一些研究者也开始考虑将时序特征[14]、情感特征[15]等引入到相似度计算公式中。

这些算法虽然在一定程度上都提高了微博类短文本的相似度计算效率,但仍然存在缺陷,如针对文本内容特征改进的相似度算法,虽然考虑了文本的语义信息,但是在处理信息量少、内容稀疏的微博文本时,普遍存在着准确性低,时间、空间消耗大的问题;引入微博非文本特征的相似度算法一定程度扩展了微博的信息内容,但现有算法还都停留在简单网络层面,大都只是引入单一层次社会网络,缺乏对微博舆情发生过程多种关系数据的有机融合。而微博舆情形成是一个复杂的过程,若要更准确地挖掘微博舆情相似度,还需要寻找一种更为全面、有效的方法揭示微博舆情的形成过程。针对上述问题,本研究拟采用超网络(Supernetwork)的思想和方法对微博舆情主题相似度进行更深层次的研究,探究多舆情要素与舆情主题形成的内在联系,就此提出超边相似度算法并进行舆情主题挖掘。相对于传统文本挖掘方法,基于超网络分析的主题相似度研究方法可以观察整个微博的社会网络总体结构特征,分析更多的指标,为挖掘具有复杂网络特征的主题信息提供了可参照模型。

2.2 超网络分析

"超网络"最早由 Y. Sheffi[16] 和 P. Denning [17] 提出,A. Nagurney[18] 给出了超网络的明确定义,指高于而又超于现存网络的网络,它在嵌套、多层、多级和多属性方面表现出自身的优越性,被广泛应用于供应链[19]、交通[20]、金融[21]及知识管理[22]等领域中。目前,针对超网络的研究主要集中在变分不等式、超图和系统科学 3 方面的研究上[23],而针对互联网文本研究通常属于后两者范畴,本研究属于系统科学范畴。超网络的多层级属性可以很好地描述网络间的作用关系,一些学者已经尝试将超网络方法应用到微博舆情研究中,如尚艳超等[24]构建了话题和用户两个维度超网络模型,潘芳等[25]在该基础上进一步考虑了网络社群舆情传播网络和社会网络之间的关系,构建了微博舆情反腐超网络模型。这些已有的超网络模型包含的特征信息过少,不足以揭示微博舆情主题发生过程,且对各层子网的揭示深度不够。笔者在前序研究中已经基于舆情传播要素构建了包含 4 层子网的微博舆情主题发现超网络模型[26],本研究在该基础上深入分析超网络模型同质节点与异质节点间的关联关系,设计一套超边相似度算法(SuperEdgeSimilarity),是对现有超网络方法中超边分析方法研究的有益补充。

3 超边相似度算法

3.1 微博舆情主题发现超网络模型建模

现实社会事件借助微博平台的关注与转发机制建立用户关系,实现信息(关键词、情感)的分享、传播和交流,形成微博舆情事件。微博舆情的发生过程,类同于现实社会的突发事件,需要厘清 5W1H(When、Where、Who、Why、How)六要素的关系[27]。一条微博信息是一个用户在情感要素和外部环境信息驱动下发布的关键词,而一个舆情事件由多条微博信息传播组成。由此,微博舆情形成关联的实体包括微博用户(Who)、时序环境(When)外驱动力、情感(How)内驱动力和关键词(When)4 类。在此基础上构建微博舆情主题发现超网络模型,包含的 4 层子网,分别是"社交子网""时序子网""情感子网"和"关键词子网"。

(1)社交子网 A(Social Network)表示参与微博话题讨论的微博用户之间的转发关系。节点是微博信息发布的用户,以微博信息的转发关系构造无向边。

(2)时序子网 T(Timing Network)表示微博舆情演化的时序阶段,本文参照生命周期理论将微博舆情演化划分为 4 个阶段,即"潜伏期→发生期→持续期→恢复期"[3]。节点表示微博舆情信息的演化阶段,相邻演化阶段存在着转化关联关系。

(3)情感子网 S(Sentiment Network)表示舆情爆发时所蕴含的情感信息,不同情感间存在着转化关联关系。本研究情感子网包含 3 个节点,分别是积极情感节点、消极情感节点和中性情感节点。

(4)关键词子网 K(Keyword Network)由微博文本的关键词构成,关键词节点之间的连线表示这两个关键词在同条微博中出现。

微博舆情超网络模型中 4 层子网之间通过超边(SuperEdge,SE)连接,$SE = \{a_i, t_m, s_n, k_j\}$,表示用户 a_i 在时序 t_m 外作用力和情感 s_n 内驱动力作用下,发表了关键词 k_j。一条超边即表示为一条微博,此处定义一条超边(微博)包含一个用户信息,一个情感信息、

梁晓贺，田儒雅，吴蕾，等. 基于超网络的微博相似度及其在微博舆情主题发现中的应用[J]. 图书情报工作，2020，64(11)：77－86.

一个时序信息和多个关键词。

3.2　超边相似度算法

微博文本存在内容短、表达随意、非规范化等特点，导致微博文本向量高维且稀疏，传统的相似度算法不能准确地度量微博短文本间的相似度。基于此，本文提出了一套基于超网络的微博相似度算法，利用超网络模型模拟微博舆情主题发生机制，依据各层子网相似度及不同子网对主题形成的贡献度构建超边相似度算法，如图 1 所示：

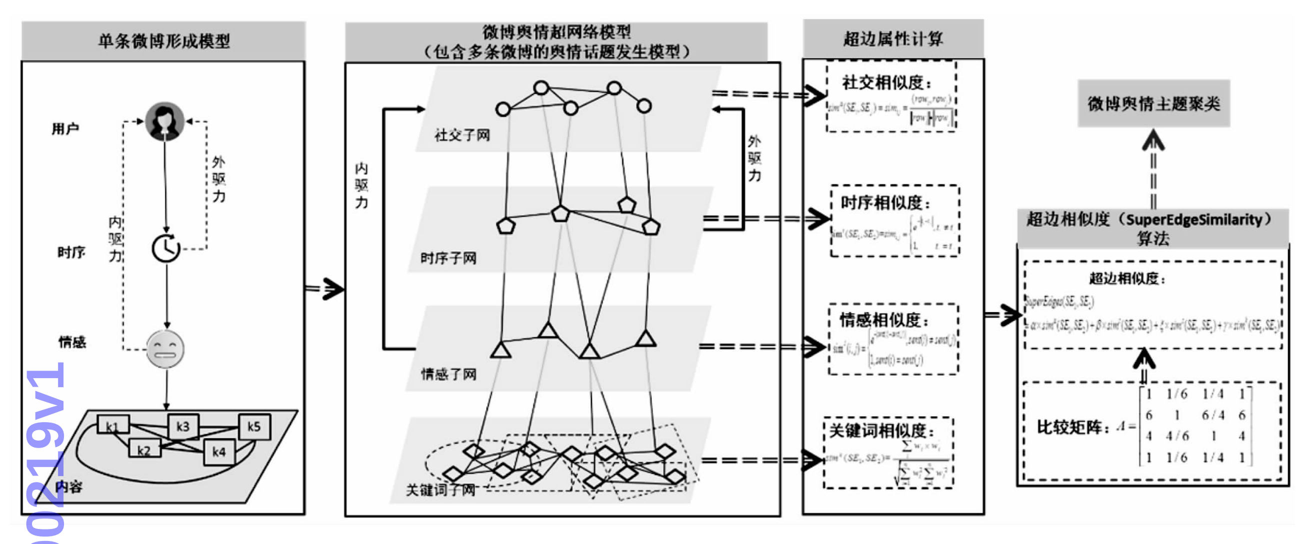


图 1　微博舆情主题发现超网络模型及超边相似度算法

本文提出的算法考虑了不同超边中包含的关键词相似程度、转发行为关系、情感转化关系和所处时序阶段的转化关系。在关键词子网中，两条超边所包含的关键词越相似，则这两条超边也越可能相似；在社交子网中，两条超边存在转发关系或转发行为越相似，其所包含的关键词也越可能相似，这两条超边也越可能相似；在时序子网中，两条超边同属于一个时序阶段或时序阶段越相近，这两条超边越可能相似；在情感子网中，两条超边包含的情感倾向相同且情感倾向相近，这两条超边越可能相似。假设微博舆情主题发现超网络模型共有 N 条超边，记为 $SE_i(1\leqslant i\leqslant N)$，假设 SE_i 和 SE_j 是待计算相似度的两条超边，由此得出以下超边相似度算法：

$$SuperEdge(SE_i,SE_j)=\alpha\times sim^{\alpha}(SE_i,SE_j)+\beta\times sim^{t}(SE_i,SE_j)+\xi\times sim^{s}(SE_i,SE_j)+\gamma\times sim^{k}(SE_i,SE_j)$$

公式(1)

其中，$sim^{\alpha}(SE_i,SE_j)$ 为超边 SE_i 和超边 SE_j 的社交相似度，$sim^{t}(SE_i,SE_j)$ 为超边 SE_i 和超边 SE_j 的时序相似度，$sim^{s}(SE_i,SE_j)$ 为超边 SE_i 和超边 SE_j 的情感相似度，$sim^{k}(SE_i,SE_j)$ 为超边 SE_i 和超边 SE_j 的关键词相似度。α，β，ξ 和 γ 分别为社交相似度、时序相似度、情感相似度和关键词相似度的权值，且满足：$\alpha+\beta+\xi+\gamma=1$，具体数值利用层次分析法确定。

3.3　超网络模型中超边属性计算

(1)社交相似度 $sim^{\alpha}(SE_i,SE_j)$。利用社交子网中用户的转发关系计算超边的社交相似度。设微博舆情主题发现超网络模型的社交子网包含 m 个节点，$p_i\in P$ $(1\leqslant i\leqslant m)$ 是社交子网节点(用户)的集合，P 中任意两个节点的相似度计算基于节点间的转发关系。参照布尔模型思想[28]，社交子网中的转发关系可以用一个矩阵 C 表示，

$$C=C_{i,j}，其中\ C_{i,j}=\begin{cases}1，节点\ i\ 与节点\ j\ 存在转发关系\\0，节点\ i\ 与节点\ j\ 没有转发关系\end{cases}$$

公式(2)

利用 $row_i=(C_{i,1},C_{i,2},\cdots,C_{i,m})(i=1,2,\cdots,m)$ 表示超边 SEi 的转发关系，则超边 SE_i 和超边 SE_j 的社交相似度计算公式为：

$$sim^{\alpha}(SE_i,SE_j)=sim_{ij}=\frac{(row_i,row_j)}{\|row_i\|\ \|row_j\|}$$

公式(3)

其中 $((row_i,row_j)=\sum_{j=1}^{m}C_{i,j}C_{i,j}$，$\|row_i\|=(\sum_{i=1}^{m}C_i^2)^{1/2})$

(2)时序相似度 $sim^{t}(SE_i,SE_j)$。微博由于其快速转发机制，使得舆情事件会在短时间内引起人们大量的转发和讨论，内容相似的微博往往在相同时间段内集中发布[29－30]。这意味着，在一个话题的发生期，人们会频繁地使用相似的关键词来进行话题讨论，而随着讨论地深入，话题会发生演化，人们讨论话题所使用

的关键词也会随之更新，但是这些更新的关键词是与演化后的话题密切相关的，所以更新的关键词彼此也是相似的。因此，相同时间段产生的关键词最可能相似，而关键词所处的阶段越相近，其产生的关键词越可能相似[31]。本文将微博舆情演化阶段(t_i)划分为潜伏期(t_1)、发生期(t_2)、持续期(t_3)和恢复期期(t4)4 个阶段。

判断完时序子网的时序阶段类型，可对不同超边所包含的时序演化关系进行度量，也就是计算时序相似度。若 SE_i 和 SE_j 时间节点处于同一个时序阶段($t_i - t_j = 0$)，则这这两条超边的时序相似度为 1，即完全相似；若 SE_i 和 SE_j 的时间节点处于不同时序阶段，则时序阶段越临近，其时序相似度越大，参照概率模型[32]思想，计算公式如下：

$$sim^t(SE_i, SE_j) = sim_{ij} = \begin{cases} e^{-|t_i - t_j|}, & t_i \neq t_j \\ 1, & t_i = t_j \end{cases}$$

公式(4)

t_i 和 t_j 为不同舆情演化阶段，其中 i 和 j 的取值范围为(1,2,3,4)，为了区分不同时序阶段的相似度差异，采用等差数据对时序阶段 t_i 进行赋值，此处综合考虑四个相似度取值的均衡性，令 $t_1 = 1$、$t_2 = 3$、$t_3 = 5$、$t_4 = 7$。

(3)情感相似度 $sim^s(SE_i, SE_j)$。由于微博舆情包含“社会脉动”和“公众情绪”[33]，情感信息是微博自媒体特征的一个体现，通常表达相似观点的微博其情感倾向也趋于一致，同理，情感趋于一致的微博更可能相似。

情感相似度计算包含如下 3 个步骤：

第一步是构建情感词典、识别超边中的情感词。通过分析微博文本情感词特征和表达习惯，总结出微博情感极性判断关键特征要素，包括情感词、表情符号、否定词和程度词，提取这类情感要素有助于准确计算超边的情感强度。基于上述分析，本研究借鉴安璐等[34]的研究成果构建了包含基础情感词典、否定词典、程度副词词典和表情符号词典的情感分析法。其中，情感基础词典选用大连理工大学提供的中文情感词汇本体库[35]；用户在微博上发布信息时常习惯附加感情符号表达情感，分析表情符号的情感极性可以辅助情感分析，本文参照大连理工大学对情感词的打分情况对微博平台上自定义的 84 个情感符号进行极性判断并逐一打分，具体参照表 1：

表 1 表情符号词典(部分)

表情符号	极性	强度	表情符号	极性	强度
[抓狂]	2	9	[赞]	1	7
[鄙视]	2	9	[笑 cry]	1	7
[怒]	2	7	[嘻嘻]	1	7
[吐]	2	7	[haha]	1	5
[哼]	2	5	[good]	1	5
[黑线]	2	5	[加油]	1	5
[哈欠]	2	3	[可爱]	1	3
[二哈]	2	3	[神奇女侠]	1	3
[白眼]	2	3	[馋嘴]	1	3

通过收集微博常用的否定词构建了否定词表(见表 2)。参照 Hownet 提供的程度词，并结合微博语言特色进行调整，构建了本文的程度词典，根据其对情感词强度的调整力度分为 7 个等级，权值分别是 0.4、0.6、0.8、1、1.2、1.4 和 1.6，以 0.2 逐级递增，见表 3。

表 2 否定词表(部分)

没	否	甭	不	别	勿	未	无
不曾	不要	未必	不太	尚未	毫不	不至于	绝非

表 3 程度副词表(部分)

权值	程度词
0.4	略、稍、稍微、有些、略为
0.6	较、蛮、一点儿、略加、或多或少
0.8	挺、越、颇、越发、愈加、相当
1	没有程度副词
1.2	那么、不少、更为、何止
1.4	实在、很、特、太、忒、更加
1.6	极、极度、格外、尤其、特别、非常

第二步进行超边情感强度计算。根据构建的情感词典，识别超边中的情感特征词极性、强度，表情符号极性、强度、否定词的个数和程度副词的调整强度。借鉴唐晓波[36]等构建的情感元组的思想，此处采用情感特征元组表示每条超边的情感特征，S = {情感极性、强度；表情符号极性、强度；否定词个数；程度副词调整强度}，所有情感元组元素均不是情感元组必备元素，即存在超边的情感元组为空的情况。对每条超边构建特征情感元组，超边的情感强度计算公式为：

$$sent(i) = \begin{cases} \dfrac{(-1)^k \times \prod_{p=1}^{p=m} wei_p(adv) \times \sum_{j=1}^{j=n} s(w_j)}{n}, & S \neq \emptyset \\ 0, \ S = \emptyset \end{cases}$$

公式(5)

其中，$sent(i)$ 为超边 i 的情感强度，情感元组为空时，即不存在情感词，此时超边情感强度记为 0。$s(wi)$

为参照基础情感词典和符号词典计算的情感强度，这里只考虑贬义、褒义和中性 3 个极性，贬义词强度设为：－1、－3、－5、－7 和 －9，褒义词情感强度设为 1、3、5、7 和 9，中性词情感感强度为 0；$\sum_{j=1}^{j=n} s(w_j)$ 为超边 i 中全部情感词和情感符号词的情感强度之和，n 为超边中情感词和情感符号的总个数；$wei(adv)$ 为情感词前后不超过 3 个词范围内的程度副词，$wei_p(adv)$ 为超边中 i 程度副词 p 的情感调整强度；$\prod_{p=1}^{p=m} wei_p(ddv)$ 表示超边中 i 全部 m 个程度副词情感调整强度的连乘积；$\sum_{j=1}^{j=n} s(w_j)$ 为微博 i 中全部情感词＋符号词的情感强度；k 为超边 i 中否定词的个数。

第三步是计算超边的情感相似度。由第二步可以获得超边的情感强度，本文情感强度数值前面的符号(正号、负号或 0)来表示每条超边可能存在的 3 种情感极性，即 $sent(i) > 0$，说明超边蕴含着积极的情感；$sent(i) < 0$，说明超边蕴含着消极的情感；$sent(i) = 0$，说明超边蕴含的情感是中立的。判断完超边的情感极性和情感强度，可进行情感相似度计算，令 $sent(i)$ 和 $sent(j)$ 表示任意两条超边的情感强度，二者差值越小则两条超边的情感相似度越大，反之，若二者差值越大则两条超边的情感相似度越小。将超边 SE_i 和超边 SE_j 情感相似度，记为 $sim^s(SE_1, SE_2)$，则

$$sim^s(i,j) = \begin{cases} e^{-|sent(i) - sent(j)|}, sent(i) \neq sent(j) \\ I, sent(i) = sent(j) \end{cases}$$

公式(6)

(4)关键词相似度 $sim^k(SE_i, SE_j)$。关键词相似度即为传统相似度算法的度量对象，本文分别选择经典的向量空间模型[32]表示关键词子网的微博文本特征、TF-IDF 方法计算关键词权重[37]、余弦相似度算法[38]作为关键词相似度度量方法。将 SE_1 和 SE_2 映射到 n 维向量空间中，可表示为 $SE_1 = (w_1, w_2, \cdots, w_n)$ 和 $SE_2 = (w'1, w'2, \cdots, w'n)$，基于词频特征的超边关键词相似度为：

$$sim^k(SE_1, SE_2) = \frac{\sum_{i}^{n} w_i \times w_i'}{\sqrt{\sum_{i=1}^{n} w_i^2 \sum_{i=1}^{n} w_i'^2}}$$

公式(7)

式中，$w_i = \mathrm{tf}_{T_i} \times \mathrm{idf}_{T_i}$，$\mathrm{tf}_{T_i}$ 为关键词 T_i 在 SE_1 中出现的次数(即 TF 值)；$\mathrm{idf}_{T_i} = \lg(N/n)$，N 为所有超边数，$n$ 为关键词 T_i 在所有超边中出现的总次数。

3.4 基于层次分析法的特征权值计算

层次分析法(The analytic hierarchy process，简称 AHP)[39]可以有效分解目标问题，从不同层次进行分析比较，实现定性与定量结合分析的决策方法。本文利用层次算法计算超边相似度不同要素的特征权值，分成如下 4 个步骤：

(1)构建层次结构模型。通过深入分析微博舆情主题的发生机制，分解成 2 层层次结构体系，据此构建微博舆情超边相似度层次结构模型，如图 2 所示：

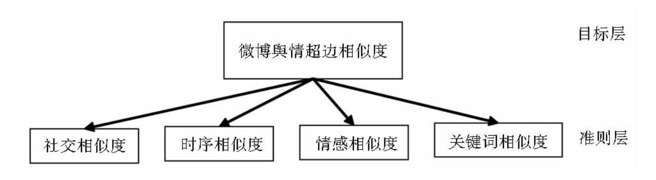


图 2 超边相似度层次结构模型

(2)构建比较矩阵。分析微博舆情主题多特征要素，关键词特征是对微博文本内容的揭示，是微博舆情主题发现的主要分析对象，所以对其赋予较高权重；情感特征作为文本内容揭示的一部分，属于次重要特征；社交特征和时序特征从侧面影响微博舆情主题的形成，且较前两者较弱，并列排在第三位。据此构建的比较矩阵如表 4 所示：

表 4 比较矩阵

相似性	社交相似性	关键词相似性	情感相似性	时序相似性
社交相似性	1	1/6	1/4	1
时序相似性	1	1/6	1/4	1
情感相似性	1	4/6	1	4
关键词相似性	6	1	6/4	6

(3)计算相对权重。依据特征值与特征向量的计算公式：$AW = \lambda_{max} W$，计算得出比较矩阵的特征向量值 $W = [0.083, 0.083, 0.333, 0.500]$，最大特征根 $\lambda_{max} = 4$。

(4)一致性检测。综合考虑一致性指标(CI)和随机一致性指标(RI)双指标分析，本文构建比较矩阵通过一致性检测。所以比较矩阵 A 所对应的特征向量可以作为权值，可得到权值向量 $W = [0.083, 0.500, 0.33, 0.083]$，即基于社交特征的微博相似度权值 $\alpha = 0.083$，基于时序特征的微博相似度权值 $\beta = 0.083$，基于情感特征的微博相似度权值 $\gamma = 0.333$，基于关键词特征的微博相似度权值 $\xi = 0.500$。

4 实验与分析

由于相似度算法的数值具有主观性，为了体现相似度方法的具体效率，本文将超边相似度计算方法应用于聚类问题。通过观察聚类结果来衡量相似度计算效果。具体操作流程参照图 3 所示：

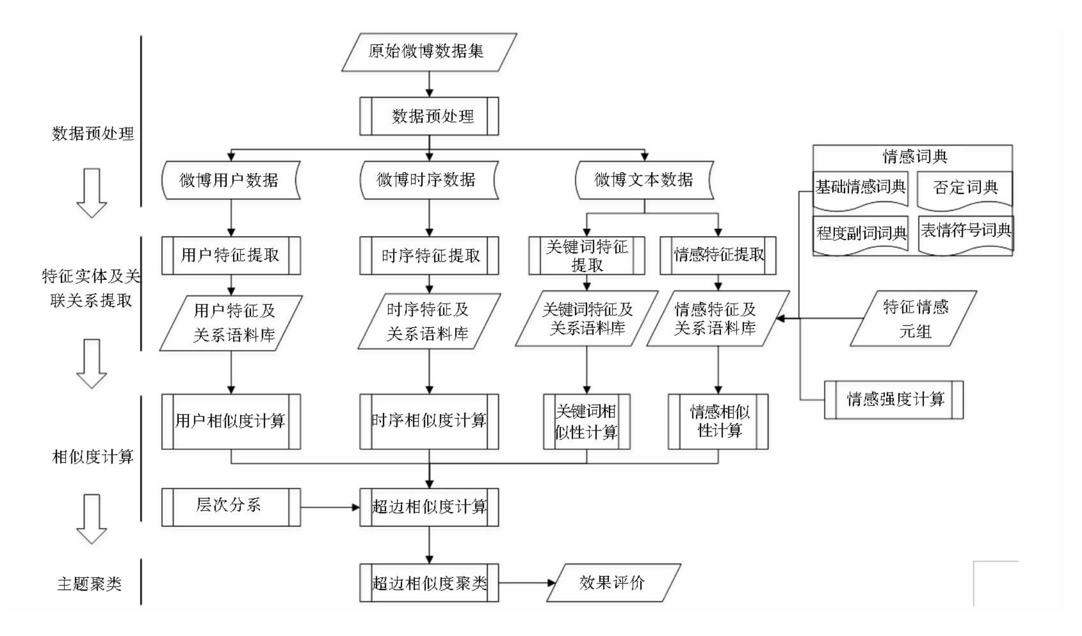


图 3 基于超边相似度算法的微博舆情主题聚类分析流程

4.1 实验数据说明

本研究的数据集为新浪微博平台，以“无籽葡萄and 避孕药”为关键词，监测时间为 2016 年 8 月 27 日 –2016 年 9 月 15 日，获取“无籽葡萄系打避孕药”这个谣言所包含的舆情信息，具体包括 ID 号、文本内容、发布时间和发布用户、转发微博账号、转发微博内容。剔除明显无关微博和相同微博（同用户名、同 ID 号、同时间的微博）；去掉停用词、超链接（URL）和一些无关符号（‘#’、‘【】’）；对于转发微博，去掉@ 姓名，将转发微博内容提前。初步清洗后得到 3 889 条数据，话题参与者共 3 600 人。由信息管理学研究人员依据网络舆情事件传播规律人工总结出“无籽葡萄系打避孕药”微博事件的主旨，并以此为依据对微博数据集进行主题类别标注。依此得到，所涉舆情事件包含了谣言发起（#1）、加深（#2）、政府、科研部门辟谣（#3、#4）、谣言摧毁（#5）、谣言后果（#6）、谣言虚假性分析（#7）、呼吁追责造谣者（#8）8 个子主题，具体内容如下：

#1、无籽葡萄系打避孕药，不敢吃葡萄，谣言发起；

#2、水果商贩和果农对话食品披露，无子葡萄喷洒了避孕药，果农不吃；

#3、引发其他食品安全联想，对政府、社会不信任；

#4、专家辟谣，无籽葡萄是通过农业技术手段培育，使用的是一种叫做赤霉素的生长调节剂，与避孕药无关；

#5、政府辟谣，呼吁大家不要信谣言、传谣；

#6、谣言致葡萄滞销、坑害果农；

#7、从常识分析谣言的虚假性：避孕药成本高，不可能用到葡萄上，人体激素不会对植物有效；

#8、造谣者无德，呼吁相关部门加强监管、严惩造谣者。

每个子主题所包含的微博数量如表 5 所示：

表 5 微博子主题划分信息

子主题类别	#1	#2	#3	#4	#5	#6	#7	#8
微博数量（条）	410	100	882	844	358	392	95	761

4.2 实验过程

采用 Python 编程调用中国科学院的 NLPIR 分词代码对清洗后的数据进行分词处理，为了提高分词效果，本研究从搜狗词库中的农业词库中筛选部分专术语，如赤霉素、植物激素、动物技术、单倍体育种等添加至用户词典导入分词系统。

对时序特征提取前，先要确定舆情事件时序演化阶段的具体时间节点。统计新浪微博中所涉“无籽葡萄系打避孕药”舆情事件的微博发布数量变化（见图 4），本文对所研究的舆情案例进行舆情传播周期的切分，周期分为 4 个阶段：潜伏期（t_1，8 月 27 日到 9 月 3）、发生期（t_2，9 月 4 日至 9 月 5 日）、持续期（t_3，9 月 6 日至 9 月 8 日）、恢复期（t_4，9 月 9 日至 9 月 14 日）。在划分阶段中，8 月 27 日到 9 月 3 日传播略有波动，但环比增长率基本保持不变，是潜伏期的主要特征；9 月 4 日当日传播量出现激增现象，并到 9 月 5 日达到了传播的最高峰值，是发生期显著的传播特征；9 月 6 日到

9月8日传播量呈现下降趋势，且增长率保持稳定，是持续期显著的传播特征；9月9日至9月14日传播量与环比增长率均波动变化，但总体传播量普遍偏低，属于恢复期特征。

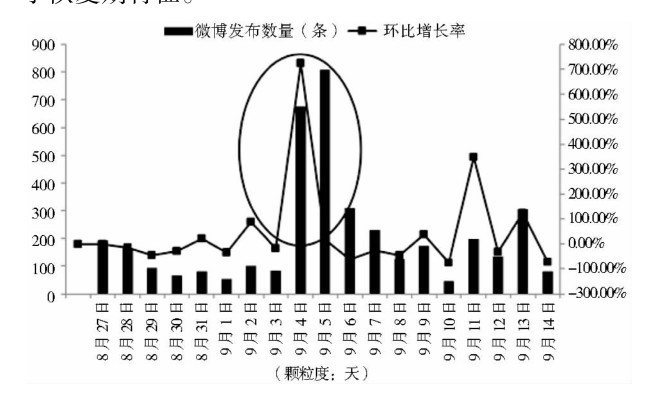


图4　新浪微博中所涉“无籽葡萄系打避孕药”事件微博发布数量统计

本研究选取 *k-means*[40] 算法进行聚类分析，*k-means* 算法是一种简单高效的聚类算法，在文本聚类中得到了广泛应用。但是 *k-means* 算法需要预先选定 k 值(聚类个数)，其聚类效果很大程度上受制于最佳k值得选择，为了消除最佳 k 值选择对实验结果的影响，本文采用手肘法和轮廓系数相结合的方法确定最佳 k 值，通过手肘法获取“肘点”取值区间，选取“肘点”取值区间内轮廓系数最大的数值作为 k 值[41]。

4.3　实验评估指标设计

实验结果的评估可以用于检验提出的方法的准确性和有效性。采用查准率 *precision*(简记为 P)、召回率 *recall*(简记为 R)和综合两者的指标 *F-measure* 值作为检验实验效果的评价指标，查准率用于检验模型的准确性，查全率用于检验模型的完备性，查准率和查全率是相互制约的关系，因此用 F 值综合评价两者。其计算公式分别如下所示：

$$p = \frac{a}{b} \qquad \text{公式(8)}$$

$$R = \frac{a}{c} \qquad \text{公式(9)}$$

$$F = \frac{2 \times P \times P}{P + R} \qquad \text{公式(10)}$$

其中，a 为实验识别正确的聚类的微博数，b 为实验识别的该类别的微博总数，c 表示微博数据集中该类的微博数目。

在对比试验分析中引入效果改善率指标(Effect Improvement，简记为 EI)，用以评价新算法的改善效率，计算公式如下：

$$EI = \frac{F_{新} - F_{旧}}{F_{旧}} \times 100\% \qquad \text{公式(11)}$$

其中，$F_{新}$ 为实验中改进算法的 F 值，$F_{旧}$ 为实验中对比算法的 F 值。

4.4　实验结果与分析

采用4.2介绍的方法确定关键词相似度算法的最优 k 值与超边相似度算法的最优 k 值相同，均为8，与人工聚类个数一致。

为了验证本文提出的超边相似度算法应用于微博文本主题聚类的效果，分别以仅考虑微博文本内容的余弦相似度聚类方法和常用的短文本聚类方法 FIHC[42] 作为对照方法。分别采用三种算法对“无籽葡萄系打避孕药”微博数据进行聚类，8个子主题的聚类效果如图5所示：

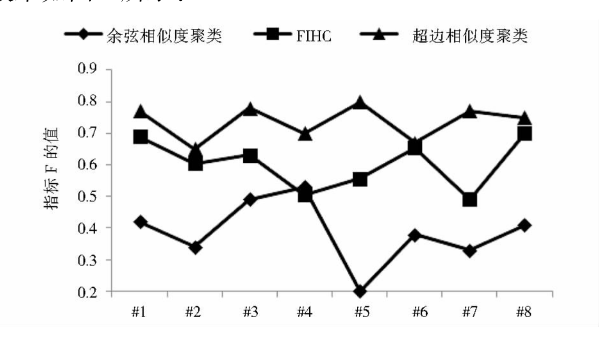


图5　不同相似度算法的聚类效果比较

采用本文提出的超边相似度聚类方法获得的8个子主题聚类指标 F 值均高于其它2种方法，即本文提出的改进算法主题识别效果普遍高于余弦相似度聚类方法和 FIHC 方法。特别是将余弦相似度聚类方法应用于微博文本时，其聚类效果普遍偏低，除去#4子主题，其余子主题的F值均在0.5以下。

纵观不同子类聚类结果，#5子主题的超边相似度算法(F 值为0.80)的聚类效果显著优于余弦相似度聚类算法(F 值等于0.20，效果改善指标为300%)和 FIHC 算法(F 值等于0.56，效果改善指标为42.86%)的聚类效果。进一步对三种聚类算法对#5子主题的聚类结果进行分析(见图6)，余弦相似度聚类方法和 FIHC 聚类方法对#5子主题聚类效果低的原因是混入了部分#1、#2、#4和#7子主题。从主题内容看，#1和#2子主题发生在舆情事件的潜伏期，这个阶段人们开始对“无子葡萄系打避孕药”这个舆情事件进行关注并展开传播，#4、#5和#7子主题发多生在舆情事件的发生期和持续期，其大都围绕“无子葡萄系打避孕药”是

谣言这个核心主题展开的不同角度的讨论,这几类子话题文字表述具有极强的相似性,所以使得仅以微博文本为分析对象的余弦相似度聚类方法和 FIHC 算法很难准确区分其相似度差异;超边相似度算法,基本识别出了#5 子主题,但混淆了少量的#4 和#7 子主题,对#4、#5 和#7 进行进一步分析,发现#4 和#5 是官方通过专业知识辟谣,#7 是大众通过常识和现有知识进行辟谣,这类微博的关键词特征、时序阶段、情感特征和社交特征都更容易趋于一致,所以超边相似度算法对这些子主题进行辨识时发生了偏差。

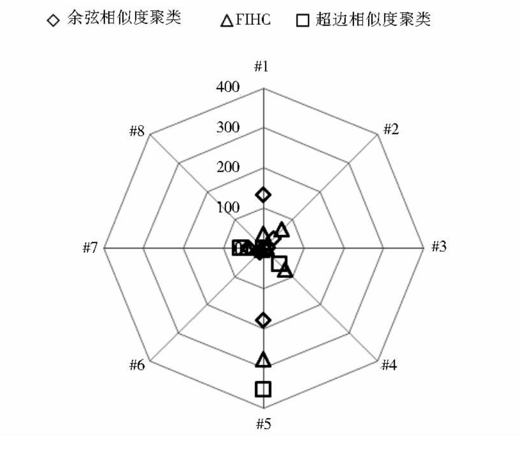


图 6 #5 子主题三种算法的聚类结果

利用余弦相似度聚类算法与 FIHC 算法对整个"无籽葡萄打避孕药"这一舆情事件的主题聚类 *F* 值分别为 0.44 和 0.50,采用超边相似度聚类方法的 *F* 值为 0.74,效果改善显著,效果改善率分别为 68.18% 和 48%。本研究从实践上进一步验证了对于微博的相似度计算仅从文本层考虑是不够的,一个舆情事件中,微博的社交转发信息、时序阶段信息和情感信息都与微博舆情主题形成密切相关,对这些特征信息的有效挖掘可以提高微博主题发现的准确度。

4.5 结论与讨论

本研究使用超网络方法模拟微博舆情主题形成传播机制,提取了与舆情主题形成密切相关的 4 个特征要素,即微博用户(Who)、时序阶段(When)、情感特征(How)和微博内容(What),构建了包含社交相似度、时序相似度、情感相似度和关键词相似度的超边相似度算法,之后将其作用到微博文本聚类中。最后通过从新浪微博采集"无籽葡萄系打避孕药"这一舆情事件数据进行试验,从 *F* 值和效果改善率这 2 方面进行评估,验证了超边相似度算法有效性。其研究结果为突发事件管理部门等利益相关者准确获取舆情主题信息、进行风险控制提供借鉴。

下一步工作主要从以下四个方面展开:①深入分析舆情主题传播特征,细化网络舆情超网络模型中的节点及关系,提高超网络模型的适用性、完整性和有效性。本文采用的超网络模型中情感子网和时序子网划分较为粗犷与社交子网和关键词子网节点数量相差悬殊,进行网络结构分析时会一定程度影响分析结果,此外粗粒度的子网结构可能会漏掉一些关键特征,使得最终结果有别于真实情况。②进一步优化各层子网相似度计算方法,提高聚类效果。特别是关键词相似度算法,本研究采用的是基于统计的 TF-IDF 方法提取微博特征词,采用余弦相似度算法进行相似度计算,缺乏语义关联,使得对于社交转发信息、时序阶段信息和情感信息都相似的微博缺乏主题辨识能力(#4、#5 和#7),后续在关键词子网也将考虑结合语义相似度的相关算法模型;③进一步拓展本研究中提出的方法在更大范围内、多网络舆情事件数据集上的实验,充分验证方法的普适性和可迁移性。④利用超网络模型分析舆情主题形成、变化的诱因,揭示舆情主题传播模式、规律,为舆情治理研究提供帮助。

参考文献:

[1] 李纲,徐伟,王馨平. 基于事件要素的组合模型微博热点事件摘要提取[J]. 图书情报工作,2018,62(1):96 - 105.

[2] 梁晓贺,田儒雅,吴蕾,等. 微博主题发现研究方法述评[J]. 图书情报工作,2017,61(17):41 - 48.

[3] 廖海涵,王曰芬,关鹏. 微博舆情传播周期中不同传播者的主题挖掘与观点识别[J]. 图书情报工作,2018,62(19):77 - 85.

[4] 刘小敏,王昊,李心蕾,等. 不同特征粒度在微博短文本分类中作用的比较研究[J]. 情报科学,2018,36(12):126 - 133.

[5] 彭敏,黄佳佳,朱佳晖. 基于频繁项集的海量短文本聚类与主题抽取[J]. 计算机研究与发展,2015,52(9):1941 - 1953.

[6] 崔金栋,孙遥遥,王欣,等. 基于 Folksonmy 和本体融合的微博信息推荐方法研究[J]. 情报科学,2015,33(10):27 - 31.

[7] ISLAM A, NKPEN D. Semantic text similarity using corpus-based word similarity and string similarity[J]. ACM transactions on knowledge discovery from data,2008,2(2):1 - 235.

[8] MA H, DI L, ZENG X, et al. Short text feature extension based on improved frequent term sets[M]. New York: Springer International Publishing, 2016:169 - 178.

[9] WEN H, WANG Z, WANG H, et al. Short text understanding through lexical - semantic analysis[C]// Proceedings of the 31st IEEE international conference on data engineering. Seoul: IEEE Computer Society, 2015:495 - 506.

[10] 黄贤英,陈红阳,刘英涛. 短文本相似度研究及其在微博话题检测中的应用[J]. 计算机工程与设计,2015,36(11):3128 -

3133.
[11] 李吉,黄微,郭苏琳. 一种基于相似度和信任度融合的微博内容推荐方法[J]. 图书情报工作,2018,62(11):112－119.
[12] KRISHNAMURTHY B, GIL P, ARLITT M. A few chirps about twitter[C]//WOSP'08 Proceedings of the first workshop on online social networks. Seattle: Association for Computing Machinery, 2008:19－24.
[13] 逯鹏,张姗姗,高庆一. 基于共同邻居的点权有限 BBV 模型研究[J]. 计算机科学,2014,41(4):49－52.
[14] 闫光辉,赵红运,任亚缙,等. 基于时间特性的微博热门话题检测算法研究[J]. 计算机应用研究,2014,31(1):43－46.
[15] 吴方照,王丙坤,黄永峰. 基于文本和社交语境的微博数据情感分类[J]. 清华大学学报(自然科学版),2014,54(10):1373－1376,1383.
[16] SHEFFI Y. Urban transportation networks: equilibrium analysis with mathematical programming methods[M]. Englewood Cliffs: Prentice-Hall,1985.
[17] DENNING P J. The science of computing: supernetworks[J]. American scientist, 1985,73(3): 127－1269.
[18] NAGURNERY A, DONG J. Supernetworks: decision-making for the information age[M]. Cheltenham: Edward Elgar Publishing, 2002.
[19] 马军,董琼,杨德礼. 时间敏感性产品供应链超网络均衡模型[J]. 系统管理学报,2015,24(4):610－616.
[20] BRICEO L, COMINETTI R, CORTES C E, et al. An integrated behavioral model of land use and transport system: a hyper-network equilibrium approach [J]. Networks and spatial economics, 2008,8(2/3):201－224.
[21] 朱莉,杜雅清. 城市群应急资源协调调配的超网络模型[J]. 数学的实践与认识,2015,45(16):27－37.
[22] 曹霞,刘国巍. 基于社会资本的产学研合作创新超网络分析[J]. 管理评论, 2013, 25(4): 115－124,157.
[23] 田儒雅,孙巍,吴蕾,等. 基于超图的图书情报领域知识合作特征分析[J]. 情报理论与实践,2016,39(10):25－30.
[24] 尚艳超,王恒山,王艳灵. 基于微博上信息传播的超网络模型[J]. 技术与创新管理,2012, 33(2):175－179.
[25] 潘芳,鲍雨亭. 基于超网络的微博反腐舆情研究[J]. 情报杂志,2014,33(8):173－177.
[26] 梁晓贺,田儒雅,吴蕾. 基于超网络的微博舆情主题挖掘方法[J]. 情报理论与实践,2017,40(10):100－105.
[27] 马宁,刘怡君. 基于超网络的舆情演化多主体建模[J]. 系统管理学报,2015,24(6):785－794,805.
[28] 张丽. 一种中文文本聚类方法的研究[D]. 哈尔滨:哈尔滨工程大学,2009.
[29] DAKKA W, GRAVANO L, IPRIROTIS P. Answering general time－sensitive queries[J]. IEEE transactions on knowledge and data engineering, 2012, 24(2): 220－350.
[30] EFORN M, LIN J, HE J, et al. Temporal feedback for tweet search with non－parametric density estimation[C]//SIGIR'14: Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval. New York: ACM Press, 2014:33－42.
[31] LIN J, EFRON M. Temporal relevance profiles for tweet search [C]// SIGIR '13: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval workshop on time－aware information access. Dublin: ACM Press, 2013. doi:10.1.1.420.611.
[32] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM,1975, 18 (11): 613－620.
[33] 李明德,蒙胜军,张宏邦. 微博舆情传播模式研究——基于过程的分析[J]. 情报杂志. 2014,33(2):120－127.
[34] 安璐,吴林. 融合主题与情感特征的突发事件微博舆情演化分析[J]. 图书情报工作,2017,61(15):120－129.
[35] 徐琳宏,林鸿飞,潘宇,等. 情感词汇本体的构造[J]. 情报学报,2008,27(2): 180－185.
[36] 唐小波,兰玉婷. 基于特征本体的微博产品评论情感分析[J]. 图书情报工作, 2016,60(16):121－127,136.
[37] 唐晓波, 房晓可. 基于文本聚类与 LDA 相融合的微博主题检索模型研究[J]. 情报理论与实践, 2013, 36(8): 85－90.
[38] 孙昌年. 基于主题模型的文本相似度计算研究与实现[D]. 合肥: 安徽大学, 2012.
[39] FAN F J, GOODMAN E D, LIU Z J. AHP (analytic hierarchy process) and computer analysis software used in tourism safety[J]. Journal of software,2013,8(12): 3114.
[40] MARQUES J P, WU Y F, et al. Pattern recognition: concepts, methods and applications [M]. Beijing: Tsinghua University Press, 2002:67－72.
[41] 王建仁,马鑫,段刚龙. 改进的 K-means 聚类 值选择方法[J]. 算机工程与应用, 2019, 55(8):1－8.
[42] CHEN C L, TSENG F S C, LIANG T. Mining fuzzy frequent itemsets for hierarchical document clustering[J]. Information processing & management, 2010, 46(2):193－211.

作者贡献说明：
梁晓贺：提出超边相似度计算方法，设计论文具体研究框架，撰写论文草稿；
田儒雅：设计各层子网相似度计算方法，修正论文研究框架；
吴蕾：负责数据处理，清洗工作；
张学福：提供论文研究思路，设计总体方案，负责论文最终版本修订。

Microblog Similarity Based on Super Network and Its Application in Microblog Public Opinion Topic Detection

Liang Xiaohe Tian Ruya Wu Lei Zhang Xuefu
Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081

Abstract: [Purpose/significance] Accurate calculation of microblog similarity can improve the efficiency of microblog topic mining, and has practical significance for public opinion governance and information security. Aiming at the problem of sparse and high-dimensional microblog text, this paper proposes a super-edge similarity algorithm incorporating non-text features of microblog. [Method/process] The mechanism of microblog public opinion was analyzed, and the formation of microblog public opinion topic formation were expressed by super network model, and the algorithm of super-edge similarity was constructed by calculating the similarity of each subnet layer and the contribution of each subnet layer to the topic formation. [Result/conclusion] It was found that the similarity method proposed in this paper is helpful to improve the topic clustering effect of microblog public opinion information. Especially for micro blog with high similarity of literal expression, it has obvious subject differentiation.

Keywords: super-edge similarity topic detection super network microblog